Applying NLP Methods to VA Electronic Notes to Improve Information Sources Available for Epidemiologic Investigation

Brett R South,MS[1], Shobha Phansalkar,RPh,MS,PhD[1], Adi V Gundlapalli,MD,PhD,MS[1], Shuying Shen,MS[1], Sylvain Delisle,MD,MBA [2], Trish Perl,MD,MSc[3], Matthew H Samore,MD[1]

[1]VA Salt Lake City Health Care system and the Department of Medicine, University of Utah, School of Medicine, Salt Lake City, UT, USA,
[2]VA Maryland Health Care System and University of Maryland, School of Medicine, Baltimore, Maryland, USA,
[3]The Johns Hopkins Hospital, Baltimore, Maryland, USA

Abstract:

We applied natural language processing (NLP) techniques to the full text of physician notes extracted from the VA's electronic medical record (VA EMR) to identify clinical and epidemiologic factors useful for investigation of flagged ILI syndrome cases.

Introduction:

•The Veterans Health Information Systems and Technology Architecture (VISTA) contains clinical and diagnostic data for all ambulatory visits and inpatient stays, including free text nursing and physician notes.

Objectives:

•Demonstrate a case finding approach to identify and screen potential cases of ILI syndrome using structured and non-structured data available from the VA EMR.

•Demonstrate clinical and epidemiologic factors useful for case investigation that can be extracted from the VA EMR using NLP methods.

Methods:

•This analysis follows a two-stage surveillance approach to identify cases for potential electronic review.

•Automated surveillance criteria were applied to the full text of electronic notes for a random sample of 15,377 patient encounters from two VA medical centers during the study period 10/01/03 to 3/31/04.

Reference Standard:

•Statistical performance of case detection models was evaluated based on manual chart review of the 15,377 encounter sample.

•ILI cases were identified based on explicit CDC criteria for ILI syndrome (Figure 1).

First Stage Surveillance:

•Various logical combinations of case detection methods based on ESSENCE and BioSense ICD-9 code sets and full note text-processing were applied to identify the most sensitive case finding approach (Table 1).

•Case detection based on text-processing methods involved coupling a negation detection algorithm called NegEx[1] adapted to VA notes with string matching for ILI concepts from the case definition mapped to the Unified Medical Language system (UMLS).

•Encounters having notes with two or more unique non-negated concepts were flagged for second stage surveillance.

Second Stage Surveillance:

•Flagged cases were electronically evaluated for case investigation variables categorized into clinical and epidemiologic factors.

•Comparisons were made between investigation variables identified by an NLP system called MedLEE[2] with ICD-9 coded variables.

•Clinical factors included fever, pneumonia, and respiratory infection as well as behavioral factors for alcohol abuse, drug abuse, and smoking status.

•Epidemiologic factors included homelessness, duration or history of illness, and previous exposure to infection that are not found elsewhere in structured data.

•Finally, similar comparisons were made between structured and non-structured data for cases admitted to the hospital for respiratory infection and pneumonia.

Figure 1. ILI case definition

1. Positive Influenza Culture or Influenza Antigen **OR**
2. The presence of any two of the following, of < 7 days duration:
   a. Cough
   b. Fever OR chills OR night sweats
   c. Pleuritic chest pain
   d. Myalgia
   e. Sore Throat
   f. Headache **AND**
3. Illness not attributable to a non-infectious etiology

Table 1. Stage 1 Surveillance ILI Syndrome

		Reference standard		Screening Method									
		Chart Review		NegEx (N)		ESSENCE (E)		BioSense (B)		E AND N		B AND N	
VA Medical Center	Cases reviewed	N	%	N	%	N	%	N	%	N	%	N	%
Baltimore	5,127	142	2.8	563	11.0	347	6.8	377	7.4	153	3.0	166	3.2
Salt Lake City	10,250	138	1.3	849	8.3	266	2.6	1017	9.9	165	1.6	226	2.2
Total	15,377	280	4.1	1,412	19.3	613	9.4	1,394	17.3	318	4.6	392	5.4

Table 2. Stage 2 Surveillance
Clinical and Epidemiologic Factors for Case Investigation

Data Source	Structured		Non-Structured	
1,412 Cases	N	%	N	%
Clinical Factors*				
Pneumonia	3	0.2	145	10.3
Fever	4	0.3	537	38.0
Respiratory infection	390	27.6	1030	72.9
Epidemiologic Factors				
Behavioral*				
ETOH abuse	36	2.5	8	0.6
Drug abuse	53	3.8	114	8.1
Smoking abuse	25	1.8	692	49.0
Demographic*				
Homeless	0	0.0	178	12.6
Exposure				
Exposure based on infectious etiology	0	0.0	1	0.1
Illness duration	0	0.0	37	2.6
History of illness	0	0.0	331	23.4
Illness duration, history of illness, or exposure	0	0.0	376	26.6

* Comparisons between ICD-9 coded data and MedLEE

Table 3. Stage 2 Surveillance
Clinical Factors for Case Investigation of Inpatient Events

Data Source	Structured		Non-Structured	
1,412 Cases	N	%	N	%
219 admissions				
Clinical Factors*				
Pneumonia	3	1.4	60	27.4
Respiratory infection	53	24.2	149	68.0

* Comparisons between ICD-9 coded data and MedLEE

Figure 3. ROC Area for ILI Case Detection Models

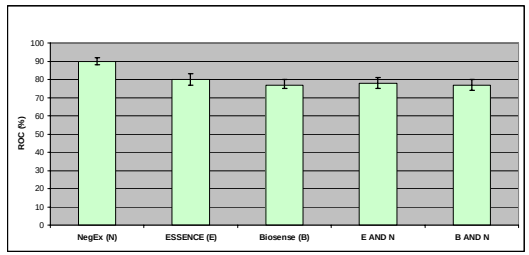

Figure 4. Sensitivity and PPV for ILI Case Detection Models

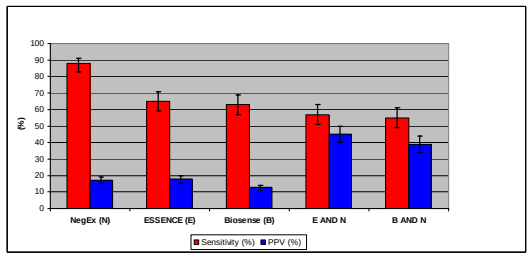

Figure 2. MedLEE XML Output: Semantic Categories and Concept Modifiers

Results:

•Manual chart review identified 280 cases of ILI used as the reference standard for case detection model performance.

First Stage Surveillance:

•NegEx produced the most sensitive case detection method and identified 1,412 potential cases of ILI (Sensitivity of 88% (83,91) (Table 1, Figure 3).

•Positive predictive values for cases having respiratory infection defined by ICD-9 code for ESSENCE AND NegEx or BioSense ICD9 code AND NegEx were 45% (40,50) and 39% (34,44) respectively (Figure 4).

Second Stage Surveillance:

•A total of 14,601 clinician notes corresponding with the 1,412 flagged cases were processed by MedLEE for second stage surveillance.

•These notes included a broad array of note types and templates.

•MedLEE maps semantic categories: (i.e. problem, procedure, bodyLoc, medications, findings) and concept modifiers: (i.e. status, certainty, time) and generates an XML output (Figure 2).

•NLP methods applied to non-structured data provide additional information that can be used further epidemiologic investigation (Table 2).

•MedLEE identified fever (38%), respiratory infection (73%), and pneumonia (10%) whereas ICD-9 codes were present for these diagnoses in <1%, 28%, and <1% of cases.

•Twenty-seven percent of all cases flagged for review had notes containing relevant ILI concepts linked to history of illness (23%), duration (3%), and exposure to infection (<1%).

•Behavioral factors were more often identified by MedLEE than by coded ICD-9 data.

•Fifteen percent of cases flagged by first stage screening were either inpatients or admitted to the hospital on the day of the randomly selected event.

•Of these cases, MedLEE identified concepts for respiratory infection in 68% of cases and concepts for pneumonia among 27% of cases.

Conclusions:

•The VA EMR is unique in that it is one of the few electronic and nearly paperless EMRs that exist in the US Health care system.

•Though our analysis focuses on syndromic surveillance, applying NLP methods to the VA EMR provides a means of extracting variables that may be useful for epidemiologic investigation that are commonly found in non-structured data elements.

•Automated case investigation could be improved by encouraging better documentation of epidemiologic factors and adding travel history to the MedLEE lexicon.

References:

1. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001 Oct;34(5):301-10.
2. Friedman C. Towards a comprehensive medical language processing system: methods and issues.Proc AMIA Annu Fall Symp. 1997;595-9.